Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2004 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

December 2004

Personalized Document Clustering: A Collaborative-Filtering-Based Approach

Chih-Ping Wei
National Sun Yat-sen University

Chin-Sheng Yang
National Sun Yat-sen University

Han-Wie Hsiao
National Sun Yat-sen University

Personalized Document Clustering: A Collaborative-Filtering-Based Approach

Chih-Ping Wei, Chin-Sheng Yang, and Han-Wei Hsiao
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan, R.O.C.
{cwei, litony}@mis.nsysu.edu.tw, hwhsiao@cc.nsysu.edu.tw

Abstract

To manage the ever-increasing volume of documents, individuals and organizations frequently organize their documents into categories that facilitate document management and subsequent information access and browsing. However, document clustering is intentional acts that reflect individual preferences with regard to the semantic coherency and relevant categorization of documents. Hence, an effective document clustering must consider individual preferences and needs to support personalization in document categorization. In this study, we design and implement a collaborative-filtering-based document-clustering (CFC) technique by incorporating an individual's and his/her neighbors' partial clusterings for supporting personalized document clustering. The empirical evaluation results suggest that the use of an individual's partial clustering can achieve a better personalized clustering result than does the content-based document clustering technique. Moreover, use of the collaborative-filtering approach for expanding an individual's partial clustering can further improve personalized clustering, measured by cluster recall and precision.

Keywords: Document clustering, Personalization, Collaborative filtering, Hierarchical agglomerative clustering (HAC), Text mining

1. Introduction

With the advances and proliferation of the Internet, information sources available on the Internet have grown tremendously in number and sheer volume, primarily because of global connectivity and ease of publishing. To facilitate individuals' information search and browsing, some emerging search engines or digital library search mechanisms (e.g., Teoma[1], vivisimo clustering engine[2], MetaCrawler[3], and WebCrawler[4]) have employed the document clustering approach to support cluster-based browsing by automatically organizing search results into meaningful categories on the fly. On the other hand, to manage the ever-increasing volume of documents generated or acquired, organizations and individuals typically organize their documents into categories (or category hierarchies) to facilitate document management and support subsequent information access and browsing. This scenario also makes document clustering an essential component for efficient and effective document management.

Essentially, document clustering is to automatically organize a large document collection into distinct groups of similar documents and to discern general themes hidden within the corpus

[1] http://www.teoma.com
[2] http://vivisimo.com
[3] http://www.metacrawler.com
[4] http://www.webcrawler.com

(Kim & Lee 2000; Kim & Lee 2002; Pantel & Lin 2002). However, document clustering is intentional acts that reflect individuals' or organizations' preferences with regard to the semantic coherency or relevant categorization of documents (Rucker & Polanco 1997). For example, given a set of research articles related to "data mining", some researchers prefer organizing by techniques under discussions (e.g., classification analysis, clustering analysis, association rules, sequential patterns), whereas others prefer categories based on application domains (e.g., banking, manufacturing, health care, telcommunications). Furthermore, even when similar clustering schemes are used, the clustering granularity may vary with different researchers. Some researchers, for example, may use a single category for all articles related to classification analysis, whereas others may employ a set of increasingly specific categories (e.g., decision tree induction, neural network, Bayes classification) for the same collection of articles. Effective document clustering therefore must consider individual preferences and needs to support personalized document categorization (Deogun & Raghavan 1986; Gordon 1991; Kim & Lee 2002).

Traditional document clustering techniques have been anchored in pure content-based analysis. As a consequence, existing document clustering techniques are not tailored to individuals' preferences and therefore are unable to facilitate personalization. Motivated by the need for personalized document clustering, this study aims to extend document clustering from content-based analysis by incorporating an individual's categorization preference into the document clustering process. Let a set of documents to be clustered be *D*. In this research, a partial clustering refers to an individual's categorization of a subset of documents in *D*. In some application environments, the partial clustering of an individual may be readily available. For example, some digital libraries or online information providers offer personal bookshelves (e.g., "my bookshelf," "my favorite," "my eNews") to users so that they can organize documents into their personal folders. When a set of documents is retrieved and should be clustered for a specific user, some of the documents in the set may have been previously organized in his or her personal folders. In this case, the partial clustering of an individual, reflecting his/her categorization preference, is available and can be employed to facilitate subsequent personalized document clustering.

However, it is possible an individual may have categorized only a small number of documents in *D*. In this case, such a small-sized partial clustering might degrade the effectiveness of personalized document clustering for this particular individual. To address the aforementioned problem, we propose the use of the collaborative-filtering recommendation approach to expand the size of an individual's partial clustering by those of other users with similar categorization preferences. Specifically, in this study, we propose a collaborative-filtering-based approach to supporting personalized document clustering and experimentally evaluate the effectiveness of our approach in comparison with a traditional content-based document-clustering technique. The remainder of the paper is organized as follows. Section 2 reviews the literature relevant to document clustering techniques and the collaborative filtering recommendation approach. Section 3 details the proposed collaborative-filtering based personalized document-clustering (referred as CFC) technique. In Section 4, we depict the experimental design and discuss important experimental results of our empirical evaluation. In Section 5, we conclude with a summary, discussion of our research contributions, and some future research directions.

2. Literature Review

In this section, we review literature on traditional content-based document clustering, a semi-supervised document-clustering technique suitable to personalized document clustering, and the collaborative filtering recommendation approach.

2.1 Content-based Document Clustering

Traditional document clustering techniques group documents on the basis of the contents of documents. The documents in the resultant cluster exhibit maximal similarity to those in the same cluster and share minimal similarity with documents in other clusters. The general process of a content-based document clustering technique consists of three main phases: feature extraction and selection, document representation, and clustering.

Feature extraction begins with the parsing of each source document to produce a set of nouns and noun phrases (commonly referred to as "features") and exclude a list of prespecified "stop words" that are non-semantic-bearing words. Subsequently, representative features are selected from the set of extracted features. Feature selection is important for clustering efficiency and effectiveness, because it not only condenses the size of the extracted feature set, but also reduces the potential biases embedded in the original (i.e., nontrimmed) feature set (Dumais et al. 1998; Roussinov & Chen 1999). Previous research commonly has employed such feature selection metrics as term frequency (TF) (which denotes the occurrence frequency of a particular term in the document collection), TF×IDF (in which IDF denotes the inverse document frequency measured by $\log(n/df)$, where n is the number of documents in the collection and df is the number of documents, including the particular term under discussion), and their hybrids (Billhardt et al. 2002; Boley et al. 1999; Larsen & Aone 1999).

According to the top-k selection method, the k features with the highest selection metric scores are selected to represent each target document in the document representation phase. Thus, each document is represented as a feature vector jointly defined by the previously selected k features. A review of prior research suggests the prevalence of the binary (which indicates the presence or absence of a feature in a document), TF, and TF×IDF (Billhardt et al. 2002; Larsen & Aone 1999; Roussinov & Chen 1999) representation methods.

In the final phase of document clustering, the target documents are grouped into distinct clusters on the basis of the selected features and their respective values in each document. Common approaches include partitioning-based (Boley et al. 1999; Larsen & Aone 1999), hierarchical (Roussinov & Chen 1999; Voorhees 1986), and Kohonen neural network (Lin et al. 1999; Roussinov & Chen 1999).

2.2 Semi-supervised Approach for Personalized Document Clustering

In addition to the described content-based document clustering approach, several prior research studies have proposed non-content-based or hybrid document clustering approaches (Yu et al. 1985; Deogun & Raghavan 1986; Kim & Lee 2000). Among them, the semi-supervised document clustering technique (Kim & Lee 2000), which considers not only content similarity but also user's perception of document similarity using a relevance-feedback mechanism, is capable of supporting personalized document clustering.

Specifically, the semi-supervised document clustering technique consists of preclustering, supervising, and reclustering phases. With the use of the hierarchical agglomerative clustering (HAC) algorithm, the preclustering phase initially places each target document in a

separate cluster and merges those two clusters whose merger produces the smallest increase in diameter. The merging process then repeats until the diameter of the merged clusters reaches a given threshold. Each of such resultant clusters is referred to as a "precluster." Subsequently, the supervising phase involves obtaining relevance feedback from a user for cluster formation in the later phase. It determines the training document set T that includes all documents within preclusters of less than η documents. Accordingly, a document d_i in T is randomly selected to serve as the query. Using this query, a set of documents in T is retrieved and presented to the user, who then judges whether each of the retrieved documents is relevant to the query (i.e., d_i). Thus, two types of document bundles are formed for d_i: positive and negative. The documents in the positive bundle, which the user has judged as relevant to d_i, are placed in the same cluster as d_i, whereas the documents in its negative bundle must be located in clusters other than d_i. Finally, the reclustering phase involves the formation of clusters for the entire document collection. The preclusters created in the first phase are assigned to the nearest positive bundle. At every precluster assignment, larger clusters are generated and the set of local cluster prototypes are incrementally updated. Finally, each residual document, which has not been retrieved or has ignored during the relevance-feedback process, is assigned to the cluster with the nearest local prototype. At this point, documents in negative bundles are examined to check whether they are located in the same clusters. If such documents are found, each of them will be reassigned to the cluster with the document's second nearest local prototype.

Although their empirical results suggest that the proposed approach outperforms a pure content-based document clustering technique (Kim & Lee 2000), the semi-supervised document clustering approach encounters several limitations or drawbacks. As described, the semi-supervised document clustering approach employs a relevance-feedback mechanism during the clustering process. However, relevance of documents to a query often depends on document traits (e.g., their quality and readability) and query intention. Thus, due to its multifacet, relevance of documents to queries may not provide appropriate estimates for measuring document similarity, possibly constraining the effectiveness of the semi-supervised document clustering approach. Moreover, the semi-supervised approach involves real-time relevance feedbacks from a user during its supervising phase. However, relevance feedbacks are time consuming and, more seriously, impractical to many document clustering applications (e.g., supporting cluster-based browsing by digital libraries and search engines), possibly limiting the applicability of the semi-supervised document clustering technique.

2.3 Collaborative-filtering Recommendation Approach

The collaborative-filtering recommendation approach identifies users whose tastes are similar to those of a target user and recommends to the target user items they have liked (Balabanovic & Shoham 1997). Several different techniques have been proposed for collaborative-filtering recommendation, including neighborhood-based, Bayesian networks, singular value decomposition with neural network classification, and induction rule learning. Among them, the neighborhood-based techniques are most prevalent (Shardanand and Maes 1995; Herlocker et al. 1999; Sarwar et al. 2000). The general process of a neighborhood-based collaborative-filtering recommendation technique encompasses two major phases: neighborhood formation and recommendation generation (Sarwar et al. 2000). The neighborhood formation phase, the model-building process for collaborative-filtering recommendation, computes the similarities between the preference of a target user and those of all other users. Several different similarity measures have been proposed (Shardanand and Maes 1995; Herlocker et al 1999; Sarwar et al. 2000), including Pearson correlation coefficient, constrained Pearson correlation coefficient, Spearman rank correlation coefficient,

cosine similarity, and mean-squared difference. For example, the similarity between a target user u_a and another user u_b using the Pearson correlation coefficient is calculated as:

$$sim(u_a, u_b) = \frac{\sum_i^m (p_{ai} - \overline{p}_a)(p_{bi} - \overline{p}_b)}{\sqrt{\sum_i^m (p_{ai} - \overline{p}_a)^2} \sqrt{\sum_i^m (p_{bi} - \overline{p}_b)^2}}$$

where P_{a_i} represents the preference score of the user u_a on item i, $\overline{p}_a$ is the average preference score of u_a, and m is the number of items co-rated by both u_a and u_b.

After the similarities between the target user and all other users are computed, the next task in the neighborhood formation phase is to form a proximity-based neighborhood with a number of like-minded users for the target user. A review of prior research suggests the prevalence of several neighborhood selection schemes that include weight thresholding (i.e., all neighbors of u_a with absolute similarities greater than a given threshold are selected) and center-based best-k neighbors (i.e., a neighborhood of a pre-specified size k is formed for u_a by simply selecting the k nearest users) (Herlocker et al. 1999; Sarwar et al. 2000).

Subsequently, in the recommendation generation phase, the preference score on a specific item j is derived for the target user based on the preferences of his/her nearest neighbors, using one of the following methods:

1. *Weighted average*: This method simply combines all the neighbors' preference scores on the item j into a prediction, using the similarities between the target user and his/her nearest neighbors as the weights (Shardanand and Maes 1995).
2. *Deviation-from-mean*: To account for preference differences in means, the deviation of a neighbor's preference score on the item j from his/her mean score is first computed, where the mean preference score is taken over all items that the neighbor has rated. Afterward, the weighted average deviation from the mean is derived across all neighbors using the similarities between the target user and his/her nearest neighbors as the weights. Finally the preference score on the item j of the target user is estimated as the sum of the target user's mean score and the weighted average deviation from the mean calculated previously (Resnick et al. 1994; Konstan et al. 1997).
3. *Z-score average*: To take into account the situation where the spread of users' preference-score distributions may be different, the z-score average method, an extension of the deviation-from-mean method, has been proposed (Herlocker et al. 1999). Neighbors' preference scores are first converted to z-scores. Accordingly, the preference score on the item j of the target user is predicted as the sum of the target user's mean score and a weighted average of the neighbors' z-scores on the item j.

3. Collaborative-filtering-based Document Clustering (CFC)

In response to the shortcomings of both content-based and semi-supervised techniques in supporting personalized document clustering, we propose a collaborative-filtering-based document-clustering (CFC) technique that incorporates a target individual's and other users' partial clusterings for estimating the categorization preference of the target individual. As shown in Figure 1, the proposed technique consists of four main phases: 1) collaborative clustering-expansion; 2) feature construction; 3) document representation; and 4) clustering.

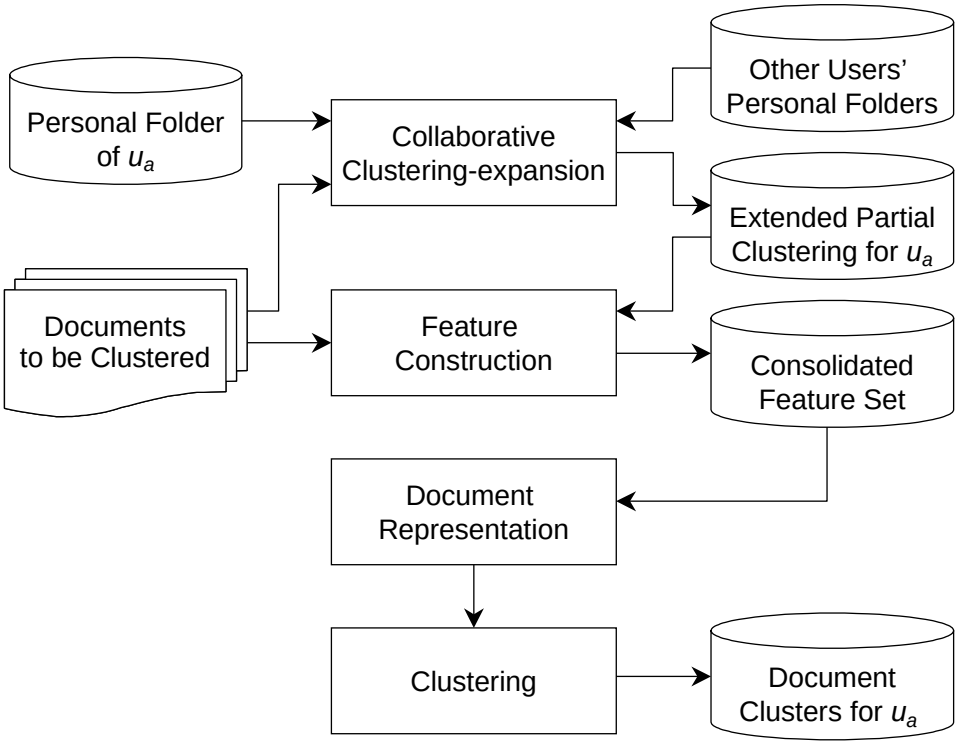

Figure 1: Process of the CFC Technique

3.1 Collaborative Clustering-expansion Phase

The purpose of the collaborative clustering-expansion phase is to expand the size of an individual's partial clustering by considering those of other users with similar categorization preferences. Two major tasks are involved in this phase: neighborhood formation and expansion of partial clustering.

To form the neighborhood for the target user u_a, we first compute the similarities between the target user and all other users based on their partial clusterings. Assume that D be the set of documents to be clustered for u_a. Let $D_{ab} \subset D$ be a subset of documents in D that exists both in the personal folders of u_a and in those of another user u_b. Furthermore, assume that C_a be the partial clustering of u_a (i.e., C_a is a set of clusters for all documents in D_{ab}, conforming to the personal folders of u_a) and C_b be the partial clustering of u_b. In this study, the similarity of the clustering preferences of u_a and u_b is estimated as a function of the similarity between C_a and C_b. Since C_a and C_b contain sets of document clusters, we adopt the concept of associations (Roussinov & Chen 1999) for measuring their similarity. Let the documents in D_{ab} can be organized in a total order and $d_i \prec d_j$ if $d_i \in D_{ab}$ appears before $d_j \in D_{ab}$ in the defined order. Hence, S_a and S_b, two sets of associations in C_a and C_b respectively, are defined as:

$S_a = \{(d_i, d_j) \mid d_i \in D_{ab}, d_j \in D_{ab}, d_i$ and d_j are in the same cluster in C_a, and $d_i \prec d_j\}$ and
$S_b = \{(d_i, d_j) \mid d_i \in D_{ab}, d_j \in D_{ab}, d_i$ and d_j are in the same cluster in C_b, and $d_i \prec d_j\}$.

Accordingly, the similarity of C_a and C_b is defined as:

$$similarity(C_a, C_b) = \begin{cases} \dfrac{2 \times |S_a \cap S_b|}{|S_a| + |S_b|} & \text{if } S_a \neq \varnothing \text{ or } S_b \neq \varnothing \\ 0 & \text{if } S_a = \varnothing \text{ and } S_b = \varnothing \end{cases}$$

Evidently, if the number of documents in D_{ab} is large, $similarity(C_a, C_b)$ would be a good

estimate of the similarity of the clustering preferences of u_a and u_b. However, a decrease of $|D_{ab}|$ would reduce our confidence on use of $similarity(C_a, C_b)$ for estimating the similarity of the clustering preferences of u_a and u_b. Taking into account the described effect of $|D_{ab}|$, we defined the similarity of the clustering preferences of u_a and u_b as:

$$similarity(u_a, u_b) = confidence(|D_{ab}|) \times similarity(C_a, C_b)$$

where $confidence(|D_{ab}|) = \begin{cases} \left(\frac{|D_{ab}|}{SigN}\right)^h & \text{if } |D_{ab}| \le SigN \\ 1 & \text{if } |D_{ab}| > SigN \end{cases}$ and $SigN$ is a pre-defined significance threshold.

After the similarities between the target user u_a and all other users are computed, we can form the neighborhood for u_a. In this study, the top n nearest users are selected and used to form the neighborhood N_a for u_a. Subsequently, the expansion of partial clustering task is undertaken to address the problem of the possibly small-sized partial clustering of u_a that might degrade the effectiveness of personalized document clustering for u_a. Let U be a subset of documents in D that exists either in the personal folders of u_a or those of any of his/her nearest neighbors in N_a. For each pair of documents d_i and d_j in U, their similarity collaboratively determined by u_a and his/her neighborhood is defined as:

$$similarity_{collaborative}(d_i, d_j) = \lambda \times f_a(d_i, d_j) + (1-\lambda) \frac{\sum_{u_b \in N_a} similarity(u_a, u_a) \times f_b(d_i, d_j)}{\sum_{u_b \in N_a} similarity(u_a, u_a)}$$

where $f_a(d_i, d_j)$ (or $f_b(d_i, d_j)$) is 1 if d_i and d_j appear in the same folder of u_a's (or u_b's) partial clustering, 0 if d_i and d_j appear in different folders, and 0.5 (i.e., unknown) otherwise. λ denotes the weight of u_a's preference-based document similarity between d_i and d_j to their overall collaborative-based document similarity.

Accordingly, based on the collaborative-based document-similarities, we perform a pre-clustering on the set of documents in U using a document clustering algorithm to obtain extended partial clusters for u_a. A hierarchical document clustering approach, specifically the HAC algorithm, is adopted in this study. A user-specified similarity threshold β is used to determine the appropriate number of clusters generated for U. Furthermore, clusters with less than δ documents are regarded as non-representative ones and, thus, are removed from the extended partial clustering EC_a of the target user u_a.

3.2 Feature Construction Phase

The purpose of the feature construction phase is to create a set of features for the target user u_a, considering not only the documents in D but also the extended partial clustering of u_a (i.e., EC_a). This phase involves three tasks, including feature extraction, selection, and consolidation.

Feature extraction converts each document in D into a set of nouns and noun phrases. We adopt the rule-based part-of-speech tagger developed by Brill (Brill 1992, 1994) to syntactically tag each word in the documents. Subsequently, we employ the approach proposed by Voutilainen (1993) to implement a noun-phrase parser for extracting noun phrases from each syntactically tagged document.

Subsequently, in the proposed CFC technique, feature selection first determines the representative features for the entire document collection D. We use TF×IDF as the feature

selection metric, due to its popularity in text categorization and document clustering research (Boley et al. 1999; Larsen & Aone 1999; Pantel & Lin 2002; Roussinov & Chen 1999). The set of top k_1 features is selected and referred to as *ALL_TF×IDF*. Moreover, because the extended partial clustering EC_a captures the categorization preference of the target user u_a, a set of features (denoted as $Partial\_\chi^2$) that best differentiates each cluster from others in EC_a is then selected using the weighted average of χ^2 statistic (Yang & Pedersen 1997) as the feature selection metric. Accordingly, the top k_2 features with the highest χ^2 statistic scores are selected and included in $Partial\_\chi^2$.

Furthermore, we consider a set of features that are frequent but potentially irrelevant to the extended partial clustering EC_a. Thus, on the basis of the TF selection metric, we select the top k_3 features (denoted as *Partial_TF*) from the documents in the extended partial clustering. The features in ($Partial\_TF - Partial\_\chi^2$) are nondiscriminative features with respect to the extended partial clustering EC_a and therefore should be excluded.

Finally, the feature consolidation task determines a set of relevant features by considering *ALL_TF×IDF*, $Partial\_\chi^2$, and $Partial\_TF$. Accordingly, the consolidated feature set employed for personalized document clustering is shown as:

$$(ALL\_TF{\times}IDF - (Partial\_TF - Partial\_\chi^2)) \cup Partial\_\chi^2)$$
$$= (ALL\_TF{\times}IDF - Partial\_TF) \cup Partial\_\chi^2).$$

As mentioned, *ALL_TF×IDF* includes k_1 features, $Partial\_\chi^2$ k_2 features, and *Partial_TF* k_3 features. Assume that approximately k features are selected for the consolidated feature set, and p% of the document collection to be clustered appears in the extended partial clustering. For the proposed CFC technique, we set $k_1 = k$, $k_2 = p\% \times k$, and $k_3 = p\% \times k$. That is, the maximal number of features in the resultant consolidated feature set is $k + (p\% \times k)$, and the minimal number is $k - (p\% \times k)$.

3.3 Document Representation Phase

Each document in the collection is represented by features of the consolidated feature set. In this study, the TF scheme was adopted as the representation method. Specifically, each document d_i is described by a feature vector $\vec{d_i}$ as:

$$\vec{d_i} = \langle v_{i1}, v_{i2}, \ldots, v_{ik'} \rangle,$$

where k' is the total number of features in the consolidated feature set, and v_{ij} is the within-document TF of the feature f_j in the document d_i.

3.4 Clustering Phase

Among the common document clustering approaches (including partitioning-based, hierarchical, and Kohonen neural network), hierarchical clustering has an advantage over partitioning-based, in that the number of clusters need not be prespecified and can be decreased (or increased) by adjusting the intercluster similarity threshold. Furthermore, the hierarchical clustering approach might achieve clustering effectiveness comparable to the Kohonen neural network (Roussinov & Chen 1999). Therefore, our proposed CFC technique employs the hierarchical clustering approach (specifically, the HAC algorithm) as its underlying clustering technique.

With the availability of the target user's extended partial clustering, some of the documents have already been grouped into clusters in the extended partial clustering. Therefore, the

HAC algorithm can use these partial clusters directly during its initial clustering stage. Specifically, the documents in each partial cluster are regarded as an initial cluster, and every document that does not appear in any partial cluster forms its own cluster. Subsequently, the two clusters with the highest intercluster similarity are merged into one cluster in the higher level in the clustering hierarchy until a termination condition (e.g., a predetermined intercluster similarity threshold) is satisfied.

In this study, the similarity of two documents d_i and d_j was estimated by the cosine similarity measure, as shown below. Furthermore, we employed the group-average link method (i.e., the average similarity among all intercluster pairs of documents) to measure the similarity between two clusters.

$$sim(d_i, d_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}| \times |\vec{d_j}|},$$

where $\vec{d_i}$ is the feature vector of the document d_i, and $|\vec{d_i}|$ is the length of $\vec{d_i}$.

4. Empirical Evaluations

This section reports the empirical evaluation of the proposed CFC technique, using a traditional content-based document-clustering technique (specifically, the HAC algorithm using the TF×IDF feature selection metric) as performance benchmarks. In the following, the evaluation design (including data collection, evaluation criteria, and evaluation procedure), parameter tuning experiments, and empirical evaluation results will be detailed.

4.1 Data Collection

The document collection for evaluation purpose consisted of 435 research articles related to information systems and technologies that were collected through keyword searches (e.g., XML, data mining, robotics) from a scientific literature digital library website (i.e., CiteSeer Scientific Literature Digital Library, http://citeseer.nj.nec.com/). For each article in the CiteSeer corpus, only the abstract and keywords were used in this evaluation study.

Furthermore, the CFC technique requires individuals' personal folders serving as partial clusterings to facilitate its collaborative clustering-expansion phase. A total of 34 subjects participated in our personal-folder collection. Because the CiteSeer corpus relates to information technology and systems, we constrained the experimental subjects to master's and doctoral students majoring in management information systems. Each subject was assigned around 50 documents randomly selected from the CiteSeer corpus and asked to manually categorize the documents without any hints. A subject could remove any document that he/she had difficulty in understanding its content or assigning it into any category. Moreover, additional 17 experimental subjects were solicited to construct their preferred clusters for the entire CiteSeer corpus (categorizing all of the 435 documents). A summary of the partial and complete clusterings generated by all experimental subjects is provided in Table 1.

Table 1: Summary of Subjects' Personal Folders and Complete Clusterings

	34 Individuals with Personal Folders			**17 Individuals with Complete Clustering**	
	Number of Documents Organized	Number of Folders Generated	Number of Documents in a Folder	Number of Folders Generated	Number of Documents in a Folder
Maximum	44	16	12	39	125
Minimum	21	5	1	10	1
Average	28.85	9.18	3.14	19.47	22.34

4.2 Evaluation Criteria

We employed cluster recall and cluster precision (Roussinov & Chen 1999), defined according to the concept of associations, to measure the effectiveness of the CFC technique and its benchmark technique. An association refers to as a pair of documents that belong to the same cluster. Assume that the clusters in the complete clustering manually produced by a subject u_a are the true categories for u_a. Accordingly, the cluster recall (*CR*) from the viewpoint of u_a is defined as:

$$CR = \frac{|CA|}{|T|}$$

where T is the set of associations in the true categories and CA is the set of correct associations that exists in both the clusters generated by the document clustering technique and the true categories. On the other hand, the cluster precision (*CP*) from the viewpoint of u_a is defined as:

$$CP = \frac{|CA|}{|G|}$$

where G is the set of associations in the clusters generated by a document clustering technique.

4.3 Evaluation Procedure

For each subject with complete clustering, we randomly took 20% of documents categorized by the subject as his/her partial clustering. Subsequently, the CiteSeer corpus was clustered by each clustering technique investigated. We measured the cluster recall and cluster precision for each technique. The overall clustering effectiveness of each technique was calculated by averaging the cluster recall and cluster precision obtained from all subjects (with complete clustering). To address the inevitable trade-offs between cluster precision and cluster recall, precision/recall trade-off (PRT) curves were employed. A PRT curve represents the effectiveness of a document clustering technique with different intercluster similarity thresholds (i.e., 0.02 to 0.98 in increments of 0.02 in this study). Evidently, as the intercluster similarity threshold increases, the average number of documents in each cluster decreases; thus, generally resulting in a higher cluster precision at the cost of cluster recall. A document clustering technique with a PRT curve closer to the upper-right corner is more desirable.

4.4 Parameters Tuning

In the tuning experiments, we randomly chose manual document clusterings from three subjects (with complete clustering) to determine appropriate values for parameters involved in each document clustering technique investigated. To obtain more reliable tuning results, we expanded the number of trials by randomly selecting 80% of the documents in the CiteSeer corpus and subsequently using this document subset for estimating the effectiveness

of each technique under a specific set of parameter values. To minimize the potential biases resulting from the sampling process, the described sampling-and-clustering process was performed ten times and the overall effectiveness for each document clustering technique was estimated by averaging the performance estimates obtained from the 10 individual sampling-and-clustering processes.

We first examined effects of the number of features (k), ranging from 100 to 500 in increments of 100, for representing documents on the effectiveness of the content-based document clustering technique. Figure 2 shows effects of different feature sizes on the clustering effectiveness of the content-based document-clustering technique. The PRT curve of the content-based technique moved in the favorable direction (i.e., getting closer to the upper-right corner) as k increased from 100 to 500. Therefore, we selected 500 as the feature size for the content-based document clustering technique.

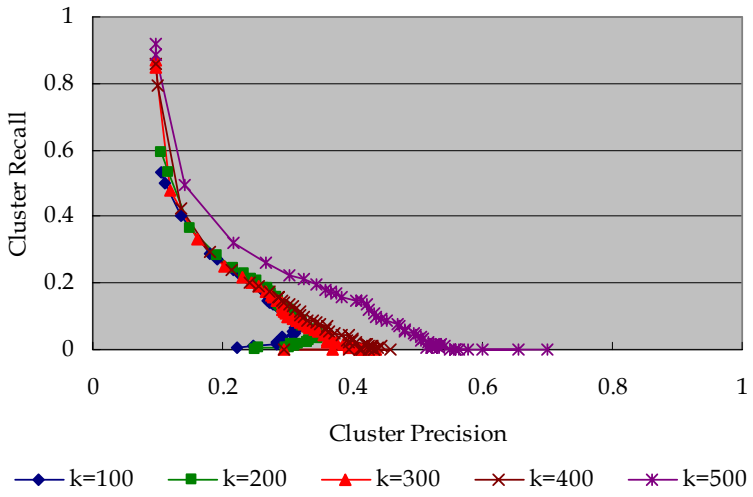

Figure 2: Effects of Feature Size k for Content-based Document-Clustering Technique

The CFC technique involves several parameters, including h and $SigN$ as required by the confidence function, n (the size of neighborhood for a target individual u_a), λ (for the collaborative-based document similarity), β (the intercluster similarity threshold for creating the extended partial clustering for u_a), δ (the size threshold for eliminating small-sized clusters from the extended partial clustering of u_a), and k (the number of features for representing documents). To reduce the magnitude of parameter tuning experiments, we set $SigN$ at 10, λ at 0.5, β at 0.5, and δ at 2 in subsequent experiments. That is, we only conducted tuning experiments for h, n, and k for the CFC technique in this study. Specifically, we investigated effects of different levels of h (i.e., 1.1, 1.3, and 1.5), n (5 and 10), and k (ranging from 100 to 500 in increments of 100) on clustering effectiveness of the CFC technique.

When tuning the parameter h, we set n as 5 and k as 500 (as with the content-based document clustering technique). Our evaluation results showed that the effects of h on the clustering effectiveness of CFC was marginal. Hence, we selected 1.3 for h. Afterward, effects of n (size of neighborhood) were examined. A better effectiveness was achieved when n = 10, as illustrated in Figure 3. Finally, we investigated effects of different feature sizes (i.e., k) and our tuning results showed that the effects of k on clustering effectiveness of the CFC technique were marginal, with k as 300 being the best. Therefore, we selected 300 as the feature size for CFC for subsequent experiments.

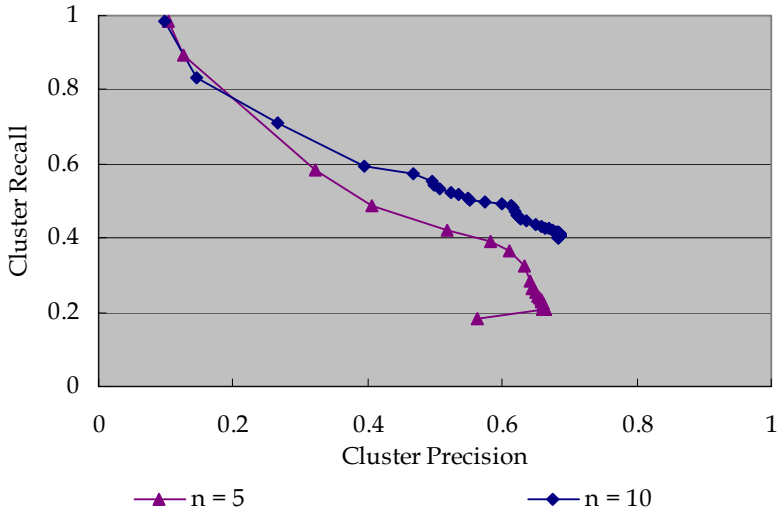

Figure 3: Effects of *n* for the CFC Technique (Using *k* = 500)

4.3 Comparative Evaluation

In the comparative evaluation experiment, the performance of the proposed CFC technique and the content-based document-clustering technique was examined. In this experiment, manual document clusterings from all 17 subjects (with complete clustering) were used for evaluation purpose. In addition, all documents in the CiteSeer corpus were included for each subject. As shown in Figure 4, the CFC technique achieved better personalized clustering results than did the content-based technique. Furthermore, we also examined the clustering effectiveness of the CFC technique when all other users' partial clusterings (i.e., *n* = 0) were not taken into account. In this case, the CFC technique is purely based on a target user's partial clustering. As also shown in Figure 4 the CFC technique with *n* = 0 still outperformed the content-based one. On the other hand, the incorporation of neighbors' partial clusterings to generate extended partial clustering for a target user had positive effects on clustering effectiveness of the CFC technique.

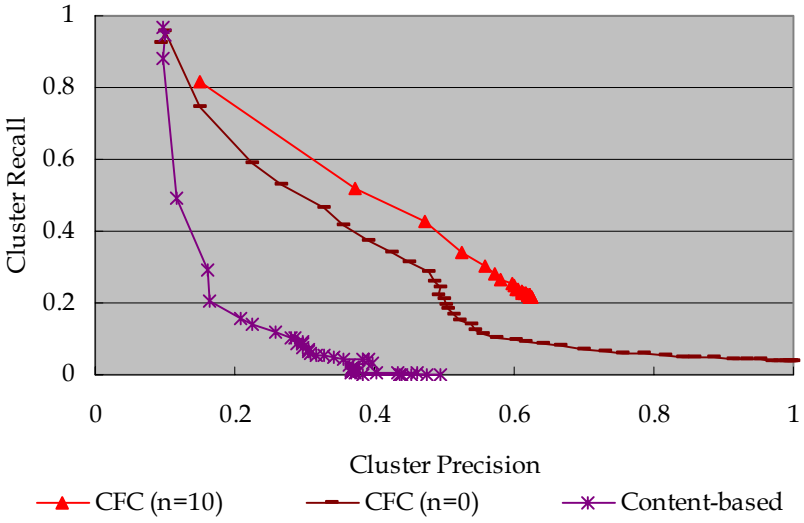

Figure 4: PRT Curves of Different Document Clustering Techniques

5. Conclusion and Feature Research Directions

Existing document clustering techniques typically generate a single set of clusters for all individuals without tailoring them to individuals' preferences and thus are unable to support personalization. Our research has been motivated by the importance of and need for personalized document clustering, especially in e-commence environments. In this study, we design and implement a collaborative-filtering-based document-clustering (CFC) technique by incorporating an individual's and his/her neighbors' partial clusterings for supporting personalization of document clustering. The empirical evaluation results suggest that the use of an individual's partial clustering can achieve better personalized clustering results than does the content-based technique. Moreover, use of the collaborative-filtering approach for expanding an individual's partial clustering can further improve personalized clustering, measured by cluster recall and precision.

Some ongoing and future research directions are briefly discussed as follows. First, our experimental study did not involve a large number of subjects for contributing personal folders and complete clusterings. A future evaluation plan involving more subjects is one of our future research directions. This research concentrated on a user's personal folders organized non-hierarchically. However, it is common that users organize their folders in a hierarchical structure. Hence, the proposed CFC technique has to be extended for accommodating users' folder hierarchies when estimating similarities of clustering preferences between users. On the other hand, the CFC technique generates a flat set of clusters. It would be desirable to extend the CFC technique to organize documents into a hierarchical cluster-structure. Finally, the empirical evaluation of this study was conducted in a laboratory setting. It would be essential to port the proposed CFC technique to a digital library and subsequently to perform empirical evaluations in such real-world setting.

Acknowledgment

This work was supported in part by the MOE Program for Promoting Academic Excellence of Universities of the Republic of China under the grant 91-H-FA08-1-4.

References

Balabanovic, M. and Shoham, Y. "Fab: Content-based, Collaborative Recommendation," *Communication of the ACM* (40:3), 1997, pp.66-72.

Billhardt, H., Borrajo, D., and Maojo, V. "A Context Vector Model for Information Retrieval," *Journal of the American Society for Information Science and Technology* (53:3), 2002, pp.236-249.

Boley, D., Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, L. "Partitioning-based Clustering for Web Document Categorization," *Decision Support Systems* (27:3), 1999, pp.329-341.

Brill, E. "A Simple Rule-based Part of Speech Tagger," *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy, 1992, pp.152-155.

Brill, E. "Some Advances in Rule-based Part of Speech Tagging," *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1994, pp.722-727.

Deogun, J. and Raghavan, V. "User-oriented Document Clustering: A Framework for Learning in Information Retrieval," *Proceedings of the 9th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1986, pp.157-163.

Dumais, S., Platt, J., Heckerman, D., and Sahami, M. "Inductive Learning Algorithms and Representations for Text Categorization," *Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management (CIKM '98),* Bethesda, MD, 1998, pp.148-155.

Gordon, M. “User-based Document Clustering by Redescribing Subject Description with a Genetic Algorithm,” *Journal of the American Society for Information Science* (42:5), 1991, pp.311-322.

Herlocker, J.L., Konstan, J.A., Borchers, A., and Riedl, J. “An Algorithmic Framework for Preforming Collaborative Filtering,” *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development of Information Retrieval*, Berkeley, CA, 1999, pp.230-237.

Kim, H. and Lee, S. “A Semi-supervised Document Clustering Technique for Information Organization,” *Proceedings of the 9th International Conference on Information and Knowledge Management*, 2000, pp.30-37.

Kim, H. and Lee, S. “An Effective Document Clustering Method Using User-adaptable Distance Metrics,” *Proceedings of the 2002 ACM Symposium on Applied Computing*, 2002, pp.16-20.

Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., and Riedl, J. “GroupLens: Applying Collaborative Filtering to Usenet News,” *Communication of the ACM* (40: 3), 1997, pp.77-87.

Larsen, B. and Aone, C. “Fast and Effective Text Mining Using Linear-time Document Clustering,” *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp.16-22.

Lin, C., Chen, H., and Nunamaker, J.F. “Verifying the Proximity and Size Hypothesis for Self-organizing Maps,” *Journal of Management Information Systems* (16:3), 1999-2000, pp.57-70.

Pantel, P. and Lin, D. “Document Clustering with Committees,” *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002, pp.199-206.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” *Proceedings of the Conference on Computer Supported Cooperative Work* (*CSCW*), Chapel Hill, NC, 1994, pp.175-186.

Roussinov, D.G. and Chen, H. “Document Clustering for Electronic Meetings: An Experimental Comparison of Two Techniques,” *Decision Support Systems* (27:1-2), 1999, pp.67-79.

Rucker, J. and Polanco, M.J. “Siteseer: Personalized Navigation for the Web,” *Communications of the ACM* (40:3), 1997, pp.73-75.

Sarwar, B.M., Karypis, G., Konstan, J.A., and Riedl, J. “Analysis of Recommendation Algorithms for E-Commerce,” *Proceedings of the 2nd ACM Conference on Electronic Commerce*, Minneapolis, MN, 2000, pp.158-167.

Shardanand, U. and Maes, P. “Social Information Filtering: Algorithms for Automating ‘Word of Mouth’,” *Proceedings of Conference on Human Factors in Computing Systems*, 1995, pp.210-217.

Voorhees, E.M. “Implementing Agglomerative Hierarchical Clustering Algorithms for Use in Document Retrieval,” *Information Processing and Management* (22:6), 1986, pp.465-476.

Voutilainen, A. “Nptool: A Detector of English Noun Phrases,” *Proceedings of Workshop on Very Large Corpora*, Columbus, Ohio, 1993, pp.48-57.

Yang, Y. and Pedersen, J.O. “A Comparative Study on Feature Selection in Text Categorization,” *Proceedings of 14th International Conference on Machine Learning*, Nashville, TN, 1997, pp.412-420.

Yu, C.T., Wang, Y.T., and Chen, C.H. “Adaptive Document Clustering,” *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Montreal, Quebec, Canada, 1985, pp.197-203.